[0012] The methods disclosed herein may comprise generating LC-MS/MS mono-allelic data for the training of allele-specific machine learning methods for epitope prediction. Such methods may comprise increasing LC-MS/MS data quality utilizing a set of quality metrics to stringently remove false positives that increases the performance of a prediction model; identifying allele-specific HLA class II binding cores from HLA-ligandome LC-MS/MS datasets; utilizing machine learning algorithms to improve HLA class II-ligand and epitope prediction; and/or identifying biological variables that impact HLA class II-ligand presentation and improve HLA class II epitope prediction, such as gene expression, cleavability, gene bias, cellular localization, and secondary structure.

[0013] Provided herein is a method comprising: (a) processing amino acid information of a plurality of candidate peptide sequences using a machine learning HLA peptide presentation prediction model to generate a plurality of presentation predictions, wherein each candidate peptide sequence of the plurality of candidate peptide sequences is encoded by a genome or exome of a subject, wherein the plurality of presentation predictions comprises an HLA presentation prediction for each of the plurality of candidate peptide sequences, wherein each HLA presentation prediction is indicative of a likelihood that one or more proteins encoded by a class II HLA allele of a cell of the subject can present a given candidate peptide sequence of the plurality of candidate peptide sequences, wherein the machine learning HLA peptide presentation prediction model is trained using training data comprising sequence information of sequences of training peptides identified by mass spectrometry to be presented by an HLA protein expressed in training cells; and (b) identifying, based at least on the plurality of presentation predictions, a peptide sequence of the plurality of peptide sequences as being presented by at least one of the one or more proteins encoded by a class II HLA allele of a cell of the subject; wherein the machine learning HLA peptide presentation prediction model has a positive predictive value (PPV) of at least 0.07 according to a presentation PPV determination method.

[0014] Provided herein is a method comprising: (a) processing amino acid information of a plurality of peptide sequences of encoded by a genome or exome of a subject using a machine learning HLA peptide binding prediction model to generate a plurality of binding predictions, wherein the plurality of binding predictions comprises an HLA binding prediction for each of the plurality of candidate peptide sequences, each binding prediction indicative of a likelihood that one or more proteins encoded by a class II HLA allele of a cell of the subject binds to a given candidate peptide sequence of the plurality of candidate peptide sequences, wherein the machine learning HLA peptide binding prediction model is trained using training data comprising sequence information of sequences of peptides identified to bind to an HLA class II protein or an HLA class II protein analog; and (b) identifying, based at least on the plurality of binding predictions, a peptide sequence of the plurality of peptide sequences that has a probability greater than a threshold binding prediction probability value of binding to at least one of the one or more proteins encoded by a class II HLA allele of a cell of the subject; wherein the machine learning HLA peptide binding prediction model has a positive predictive value (PPV) of at least 0.1 according to a binding PPV determination method.

[0015] In some embodiments, the machine learning HLA peptide presentation prediction model is trained using training data comprising sequence information of sequences of training peptides identified by mass spectrometry to be presented by an HLA protein expressed in training cells.

[0016] In some embodiments, the method comprises ranking, based on the presentation predictions, at least two peptides identified as being presented by at least one of the one or more proteins encoded by a class II HLA allele of a cell of the subject.

[0017] In some embodiments, the method comprises selecting one or more peptides of the two or more ranked peptides.

[0018] In some embodiments, the method comprises selecting one or more peptides of the plurality that were identified as being presented by at least one of the one or more proteins encoded by a class II HLA allele of a cell of the subject.

[0019] In some embodiments, the method comprises selecting one or more peptides of two or more peptides ranked based on the presentation predictions.

[0020] In some embodiments, the machine learning HLA peptide presentation prediction model has a positive predictive value (PPV) of at least 0.07 when amino acid information of a plurality of test peptide sequences are processed to generate a plurality of test presentation predictions, each test presentation prediction indicative of a likelihood that the one or more proteins encoded by a class II HLA allele of a cell of the subject can present a given test peptide sequence of the plurality of test peptide sequences, wherein the plurality of test peptide sequences comprises at least 500 test peptide sequences comprising (i) at least one hit peptide sequence identified by mass spectrometry to be presented by an HLA protein expressed in cells and (ii) at least 499 decoy peptide sequences contained within a protein encoded by a genome of an organism, wherein the organism and the subject are the same species, wherein the plurality of test peptide sequences comprises a ratio of 1:499 of the at least one hit peptide sequence to the at least 499 decoy peptide sequences and a top percentage of the plurality of test peptide sequences are predicted to be presented by the HLA protein expressed in cells by the machine learning HLA peptide presentation prediction model.

[0021] In some embodiments, the machine learning HLA peptide presentation prediction model has a positive predictive value (PPV) of at least 0.1 when amino acid information of a plurality of test peptide sequences are processed to generate a plurality of test binding predictions, each test binding prediction indicative of a likelihood that the one or more proteins encoded by a class II HLA allele of a cell of the subject binds to a given test peptide sequence of the plurality of test peptide sequences, wherein the plurality of test peptide sequences comprises at least 20 test peptide sequences comprising (i) at least one hit peptide sequence identified by mass spectrometry to be presented by an HLA protein expressed in cells and (ii) at least 19 decoy peptide sequences contained within a protein comprising at least one peptide sequence identified by mass spectrometry to be presented by an HLA protein expressed in cells, such as a single HLA protein expressed in cells (e.g., mono-allelic cells), wherein the plurality of test peptide sequences comprises a ratio of 1:19 of the at least one hit peptide sequence to the at least 19 decoy peptide sequences and a top percentage of the plurality of test peptide sequences are